WADC TECHNICAL REPORT 58-574 (II)

STUDIES IN RESEARCH METHODOLOGY:

II. Consequences of Violating Parametric Assumptions — Fact and Fallacy

James V. Bradley

Aerospace Medical Laboratory

SEPTEMBER 1959

WRIGHT AIR DEVELOPMENT CENTER
AIR RESEARCH AND DEVELOPMENT COMMAND
UNITED STATES AIR FORCE
WRIGHT-PATTERSON AIR FORCE BASE, OHIO

WADC TECHNICAL REPORT 58-574 (II)

STUDIES IN RESEARCH METHODOLOGY:

II. Consequences of Violating Parametric Assumptions — Fact and Fallacy

James V. Bradley

Aerospace Medical Laboratory

SEPTEMBER 1959

Project No. 7184
Task No. 71581

WRIGHT AIR DEVELOPMENT CENTER
AIR RESEARCH AND DEVELOPMENT COMMAND
UNITED STATES AIR FORCE
WRIGHT-PATTERSON AIR FORCE BASE, OHIO

1,000 — November 1959 — 11-325

FOREWORD

This report was prepared by the Psychology Branch of the Aerospace Medical Laboratory, Directorate of Laboratories, under Research and Development Project No. 7184, Task 71581, with Dr. John P. Hornseth acting as Task Scientist. The research reported herein was conducted by the author at Antioch College using the facilities of Contract No. AF 33(616)-3404. The author is indebted to Darwin P. Hunt, Charles A. Baker and Melvin J. Warrick for critical reviews of early drafts of the report. He especially wishes to thank Dr. R. C. Geary and the Editor and Trustees of Biometrika for permission to quote at length from Geary's "Testing for Normality" (36).

ABSTRACT

Methods of investigating the effects of assumption-violation are examined. Particular attention is given to methodological, and other, bias operating in favor of the conclusion that parametric tests are extremely insensitive to violations of their assumptions. Fallacious arguments advanced in support of this conclusion are discussed. Using a new method, the effect of nonnormality upon the probability levels and power of the critical ratio test is investigated. Results show that under certain perfectly realistic conditions the test is rendered completely powerless by the violation investigated.

PUBLICATION REVIEW

This report has been reviewed and is approved.

FOR THE COMMANDER:

Walter F. Grether

WALTER F. GRETHER
Director of Operations
Aerospace Medical Laboratory

INTRODUCTION

The effect of violations of assumptions upon the probability levels of parametric tests has been the subject of a number of investigations, both mathematical and empirical. The results of these studies, taken as a whole, indicate that under certain qualifying conditions familiar parametric tests suffer only slight loss of precision when used in mild violation of their assumptions. They also indicate that under other perfectly realistic conditions probability levels are quite appreciably distorted by moderate assumption violations. Unfortunately, a great deal of the literature emphasizes only the former effect and does so without giving equal emphasis to the conditions which qualify it. Furthermore, much of the evidence for the former effect was obtained by methods biassed in favor of finding parametric tests insensitive to departures from the assumed conditions. The result has been an apparently widespread acceptance of the unwarranted generalization that parametric statistical tests can be used with negligible loss of precision under all but the most drastic violations of their assumptions. It is not the intention of the present article to embrace the equally absurd position that parametric tests are valueless when their assumptions are in any way violated (which is practically certain to occur). However, in an effort to regain perspective, certain fallacious arguments and methods used in support of the "insensitivity doctrine" will be discussed. Also the danger of naive application of the insensitivity doctrine to time scores and errors, typical psychological measurements, will be illustrated.

FALLACIOUS BELIEFS ASSOCIATED WITH THE "INSENSITIVITY DOCTRINE"

Some curious and naive arguments have been voiced in support of the insensitivity doctrine by lay statisticians. One argument goes that unless the violations are so extreme as to be "readily detectable" in the data, the consequences of violation will not be serious. The argument, however, is a fallacious one. For a violation of any given extent in the parent population, the larger the sample the more readily detectable the violation will be, since

it will become more and more obvious that the peculiarities in the data are due to the violation rather than to sampling error. However, many violations of assumptions have their most serious consequences when samples are smallest. For example, for tests based on sample means, such as Student's t-test, the assumption of normality generally becomes less and less important with increasing sample size because of the central limit effect. However, a given "degree" of nonnormality is more "readily detectable" in large samples than in small ones.

Another argument is that extreme violations of assumptions can be ignored because they are very rare. Therefore, it is implied that since their probability of occurrence is very small it can be "lumped" with the probability of making a Type I or Type II error, whichever type of error the violation causes. A subtle fallacy, therefore, is introduced by implicitly changing the definition of Type I and Type II errors. Under the new definition, the significance level, α, becomes the proportion of incorrect rejections for the infinite number of potential different experiments, with true null hypotheses, conducted by the same experimenter. This contrasts with the usual definition in which α is the proportion of false rejections for the infinite number of replications of the same experiment with true null hypothesis.

Another odd argument is that it really doesn't matter if data whose true chance probability is .05 are found to have a "tabled" probability of .03 or .07, because the .05 level is more or less arbitrary and the .03 or .07 levels would have served just as well. This argument ignores the fact that, arbitrary or not, the .05 level represents a consistent probabilistic reference point and serves to make one experiment statistically comparable with another. Furthermore, as will be forcefully demonstrated later, departures of α from its nominal level are accompanied by changes in the power of the test and these changes may be far more critical to the usefulness of the test than are the changes in α.

A final argument is that even when the violations themselves are extreme and obvious, their effect upon probability levels is trifling because the absolute difference between true and nominal significance levels is

small. However, since the significance level itself is a small number, the relative error corresponding to a small absolute error may be quite large. Since the relative error is obviously the appropriate index of distortion, it is amazing to find this fallacy to be fairly widespread.

These arguments are seldom stated explicitly in the literature, although they are sometimes expressed orally by lay research workers who have acquired them, directly or by reputation, from publications in which they appeared to have been implied. These publications are fa. from blameless for the misapprehensions of the laity. Many of their results were acquired by methodologically dubious procedures and their conclusions were often incautiously presented.

SUBJECTIVE AND METHODOLOGICAL BIAS IN STUDIES OF THE EFFECTS OF ASSUMPTION VIOLATION

A great deal of the overemphasis upon the insensitivity of parametric tests to violations of their assumptions appears to be the result of subjective bias. When the earlier studies of the effects of assumption-violation were conducted, practically the only versatile and well developed tests available were parametric tests. Since there was no really appealing alternative to the parametric test, the investigator was motivated to find all but the most drastic departures from the test statistic's theoretical distribution to be a "good approximation" to it. This state of mind has been eloquently described by Geary (36):

> "Our historian will find a significant change of attitude about a quarter-century ago following on the brilliant work of R. A. Fisher who showed that, when universal normality could be assumed, inferences of the widest practical usefulness could be drawn from samples of any size. Prejudice in favor of normality returned in full force and interest in non-normality receded to the background (though one of the finest contributions to non-normal theory was made during the period by R. A. Fisher himself), and the importance of the underlying assumptions was

almost forgotten. Even the few workers in the field (amongst them the present writer) seemed concerned to show that 'universal non-normality doesn't matter': we so wanted to find the theory as good as it was beautiful. References (when there were any at all) in the text-books to the basic assumptions were perfunctory in the extreme. Amends might be made in the interest of the new generation of students by printing in leaded type in future editions of existing text-books and in all new text-books:

Normality is a myth; there never was, and never will be, a normal distribution.

This is an over-statement from the practical point of view, but it represents a safer initial mental attitude than any in fashion during the past two decades."

Another type of subjective bias is largely semantic although it, too, has overtones of wishful thinking. A favorite conclusion among certain investigators of assumption-violation has been that probability levels of parametric tests are not appreciably distorted by moderate departures from the test's assumptions. Presumably what the investigator means by "moderate" violation is a violation of the degree investigated, and what he means by "unappreciable" distortion of probabilities is the degree of distortion which he obtained. Since "moderate" and "appreciable" are somewhat subjective terms, this type of implicit quantitative definition would be reasonable enough if only it were clear to the reader. One must presume, however, that sometimes it is not. And if the reader naively supplies his own definitions for these terms, as he is likely to do if he reads only the abstract of the article, he is quite likely to acquire the wrong impression.

Studies of the effects of assumption-violations have generally fallen into one of three categories: mathematical, empirical, and studies based on Fisher's method of randomization. The introduction of objective or methodological bias will be discussed simultaneously with the description of these types of studies.

In the mathematical studies the formula expressing the distribution of the test statistic is recalculated and adjusted so as to give the true, or

very nearly true, distribution of the test statistic under the particular assumption-violating conditions being investigated. The true mathematical distribution obtained with the recalculated formula is then compared with the theoretical distribution which takes no account of the violation of assumptions. These studies have the tremendous advantage of giving, usually, an exact, or nearly exact, picture of the consequence of a precisely defined violation of assumptions. They suffer, however, from the fact that the types of violations investigated are more likely to be chosen for their mathematical convenience than for their prevalence in practice. While the mathematical studies are prone to be unrealistic, they are generally free from gross methodological bias. They appear to be far superior to most studies of the other two types.

In studies using the method of randomization, a sample is drawn from an assumption-violating population, and the test statistic is calculated in the usual way and referred to the theoretical probability tables. Then the "true" probability of obtaining that value of the test statistic is obtained in the following manner. The scores obtained under the various treatments are reassigned to treatment categories in all possible ways without regard to the treatment to which a score originally "belonged", but with the proviso that the original number of scores in each category is not altered. The number of such reassignments which would result in a recalculated value of the test statistic as extreme or more so than that originally obtained is divided by the total number of possible reassignments. This fraction is regarded as the "true" probability of the originally obtained value of the test statistic and is compared with the tabled or theoretical probability. The method of randomization is the basis for statistical tests which are legitimate in their own right; however, its validity as a method of determining the exact effect of assumption-violations is questionable. The true distribution of a parametric statistic whose assumptions are violated is obtained by calculating the statistic for each of an infinite number of random samples from the assumption-violating population. Calculating the statistic for all possible reshufflings of data from a single obtained sample is not the equivalent, and distributions of the test statistic obtained by these two methods will not be the same. The method of randomi-

zation, therefore, suffers from the fact that the "population" investigated is not the parent population from which the sample was drawn, but is simply the set of scores comprising the sample.

The empirical studies have the advantage that many of the assumption-violating populations used have been fairly typical of those encountered in practice. They are extremely prone, however, to methodological bias. In the empirical studies, a large number of samples is drawn from a known population which does not meet the assumptions of the statistical test. A value of the test statistic is then calculated from the data in each sample, thus giving an empirical distribution of values of the test statistic under assumption-violating conditions. The empirical distribution is then compared with the theoretical distribution of the test statistic. The comparison is complicated, however, by the fact that the empirical distribution does not portray exactly the true distribution of the test statistic under the assumption-violating conditions, but rather portrays that distribution somewhat distorted by the superimposed fluctuations due to sampling error. It is not known therefore how much of the difference between the empirical and theoretical distributions is due to violation of assumptions and how much is attributable to sampling error. Tests of "goodness of fit" have sometimes been employed to determine the chance probability of obtaining the empirical distribution as a sample from the population defined by the theoretical, tabled distribution. However, since it is known that violations of assumptions distort the distribution of a test statistic, the null hypothesis, i.e., that the theoretical distribution is the population from which the empirical distribution was drawn, is known on theoretical grounds to be false and the test of goodness of fit cannot carry its usual meaning. If the test of fit is conducted and interpreted in the usual manner, failure to reject can only mean that a Type II error has been committed. If, on the other hand, the test is conducted with the expectation that a "serious" violation of assumptions will insure rejection and that failure to reject is indicative of a violation of only "mild" consequence, its application and interpretation become an art. For any consistent test of fit the power of the test approaches one as sample size approaches infinity. Therefore, even if violation of

assumptions produced only an infinitessimal distortion in the distribution of the investigated test statistic, rejection of the hypothesis of fit would be certain if sample size were sufficiently large. On the other hand it is well known that spectacular discrepancies between fact and hypothesis are unlikely to be detected by statistical test if sample size is sufficiently small. The experimenter's art, therefore, manifests itself in selection of an "N" of just the proper magnitude to insure rejection when the fit is just poor enough to constitute a consequence of assumption-violation which he deems "serious". Thus the test of fit may serve merely to lend an appearance of objectivity to what, unknown to him, is ultimately the experimenter's opinion. In this case the probability level of the test of fit acts as an index of opinion upon which is superimposed some noise as a consequence of the indirect, and in fact circuitous, route by which it was obtained.

Besides treating inability to reject as evidence for acceptance of a null hypothesis known to be false, the goodness of fit tests have introduced methodological bias by the manner of their application. Chi-square, which has been used almost exclusively for such tests, is particularly unsuitable for the purpose because of its requirement that expected cell frequencies be large. Frequently the fit between the theoretical and empirical curves was fairly good over the central hump of the distributions, becoming increasingly poor at more and more remote tail positions (an effect, incidentally, which was seldom remarked upon by the investigator). Since the remote tail regions are, by nature, "improbable", the corresponding expected frequencies for the tail intervals were generally very small, even when the total amount of data taken was quite substantial. Therefore, in order to bring the expected frequencies up to the level required by chi-square, the cells corresponding to the tail intervals were generally pooled, thus, of course, obscuring the poorness of fit at the tails and biassing the test in favor of the conclusion that the fit was "good". Pooling the tail cells is particularly misleading. Since the tail areas constitute the rejection region, their distortions are of greater consequence than those in any other portion of the statistic's distribution in determining the seriousness of the violation of assumptions which produced them.

A NEW METHOD INVESTIGATING THE EFFECT OF NONNORMALITY, IN PSYCHOLOGICAL MEASUREMENTS, UPON THE PROBABILITY LEVELS AND POWER OF THE CRITICAL RATIO TEST

Despite the obvious presence of the effect in their data, previous investigators have seldom taken note of the rapid worsening of fit between true and theoretical distributions of the test statistic as the remote tail regions are approached. Few have investigated the effect of assumption violations upon the power, rather than upon the probability levels, of the test. The importance of these neglected considerations will be demonstrated in the investigation to follow. The study will be applied to measurements typical of research in experimental psychology, namely time scores and errors. The distributions investigated are in no way contrived, but were obtained under the conditions of a routine human-engineering experiment. They were obtained, from a single subject, however, and are not necessarily representative of multi-subject distributions. The investigation, although applied to empirical distributions, will employ a technique which obtains the "true" probability of the test statistic uncomplicated by the presence of sampling error (except in a sense to be discussed later). It therefore incorporates certain advantages which have hitherto been features exclusively enjoyed by the empirical or by the mathematical studies. (It suffers, however, from the fact that the true probability can only be obtained, and compared with the theoretical one, in a restricted range of the theoretical distribution of the test statistic, a range over which the investigator has no control.)

All absolute time or error scores violate the parametric assumption of normality. The normal distribution extends from minus infinity to plus infinity, but it is impossible to have a negative absolute time or error score (and, if there is a maximum of one error per trial, it is impossible to have more errors than trials). Furthermore, in the case of timed performance of a task, there must be some physiological limit to the speed with which the task can be performed. It is not unreasonable to suppose

that this physiological limit is approximated by the lower boundary of the time interval represented by the lowest score recorded in a very large number of trials. Thus scores lower than the physiological limit cannot be drawn from the parent population from which a time score distribution was obtained, nor can scores lower than zero be obtained from an error population. They are impossible and their cumulative, as well as point, probabilities are zero. An impossible value for a single such score must <u>a fortiori</u> be an impossible value for the mean of a sample of N scores, since all scores below it are also impossible.

If a sample of N observations with mean $\bar{X}$ has been drawn from a population assumed to be normally distributed and known to have variance σ^2, the statistic $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}}$ can be used to test the hypothesis that the population mean has the value μ. If the largest impossible value of $\bar{X}$ below the mean, the true value of μ, and the true value of σ from a time-score or error population are substituted into the above formula, the result is the largest impossible left-tailed value Z can assume at a given sample size when the null hypothesis is true. The true probability of this or a smaller value of Z is, of course, zero; the theoretical probability is obtained from the Z tables, i.e., normal tables. Since the true values of μ and σ were substituted into the formula for Z, the discrepancy between the true and theoretical probabilities can only be attributed to violation of the assumption of normality. Again, if we substitute into the formula the theoretical, i.e., tabled, value of Z corresponding to the left-tailed α level of significance, the largest impossible value of $\bar{X}$ below the true population mean and the true value of σ, solving for μ gives us the largest falsely hypothesized value for the population mean which it would be impossible to reject, when using the left-tailed α level of significance, on the basis of a sample of size N drawn from a time-score or error population. These two procedures were used to investigate the effect of nonnormality, in the time-score and error distributions about to be described, on the significance levels and power of the Z test.

Figure 1 shows a distribution of 2520 reach-and-operation-time scores for the operation of a push button by a single subject. Times were recorded

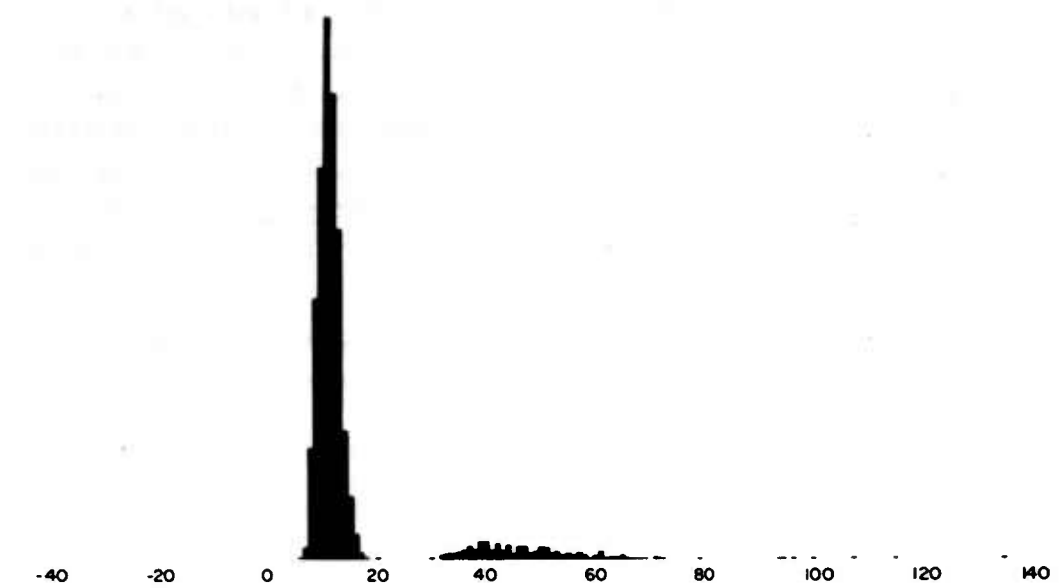

Figure 1. Frequency Distribution for Empirical Time-Score Population

to the nearest hundredth of a second. The lowest score recorded was 6, so the physiological limit and largest impossible score are taken to be 5.5. The mean and standard deviation are 15.1067 and 12.1207 respectively. The long positive tail represents time scores for trials in most of which the subject missed the push button on the first thrust of the finger, requiring multiple thrusts to operate it. Defining an error as missing on the first thrust and defining an error score as the number of errors in ten trials, Figure 3 shows the distribution of error scores in the 252 nonoverlapping blocks of ten consecutive trials for the same experiment. Error scores can range from 0 to 10; therefore, in this case the largest impossible score below the mean is a value infinitesimally smaller than zero. The mean and standard deviation of the empirical error-score distribution are 1.0357 and

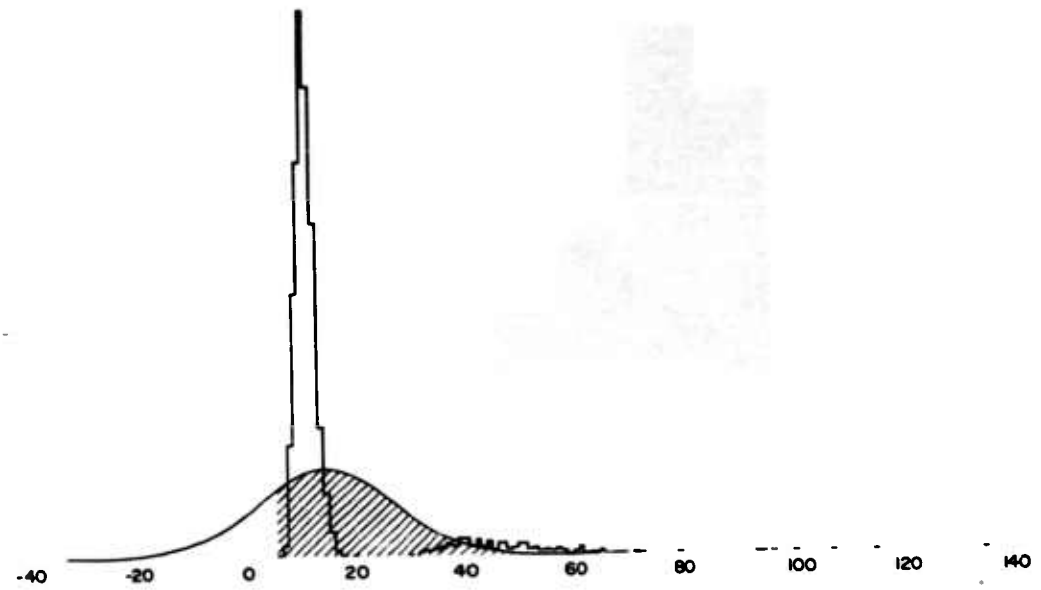

Figure 2. Time-Score Distribution (Histogram) and Normal Distribution with same Mean, Variance and Area. Unshaded area of normal distribution covers impossible scores.

.981331 respectively. The empirical distribution is quite close to the theoretical distribution of error scores based on the assumption of independence, i.e.,no sequential effects. This theoretical distribution is simply the binomial distribution of the number of errors, r, in n = 10 trials when the probability, p, of an error on a single trial is .10357. The mean is np = 1.0357 as before; but the standard deviation, $\sqrt{np\,(1-p)}$ = .963552, is slightly smaller.

The methods of analysis discussed earlier were applied to the time-score and error distributions just described. Table I takes the null hypothesis to be true and determines the largest theoretical left-tailed cumulative

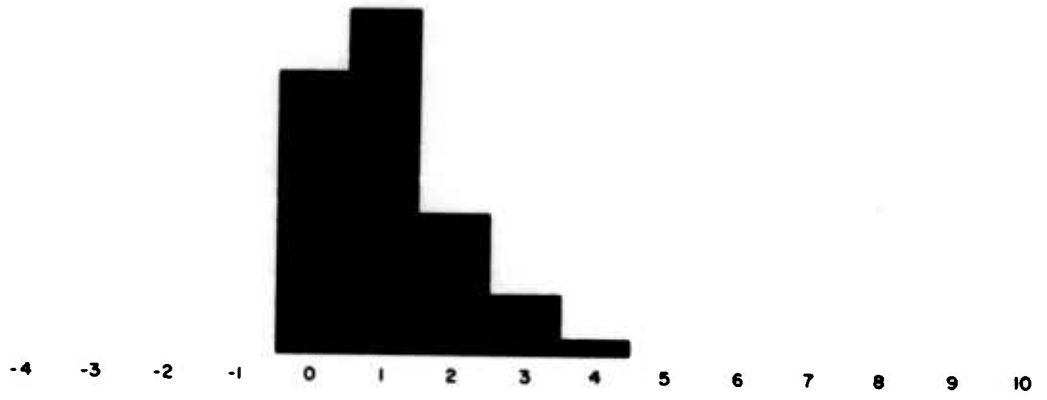

Figure 3. Frequency Distribution for Empirical Error-Score Population.

probability corresponding to a true cumulative probability of zero. It therefore shows the discrepancy, Pr (Z), between ordinates of the theoretical and true cumulative distributions of Z at the point at which the true cumulative distribution curve just begins to ascend from the x-axis. At this and lower points, the relative error between theoretical and true probabilities is infinite. Table 2, in a sense, shows the practical consequence of the effects present in Table 1. It gives the most extreme one-sided alternative hypothesis for which the Z test would have zero power to reject, i.e., for which rejection would be completely impossible.

The preceding analysis, in the case of the time scores, is strictly valid only if the obtained distribution is regarded as a population, in which case any score having zero relative frequency is impossible by definition. If, however, it is regarded merely as a very large sample, it might be objected that the scores defined earlier as impossible are merely so improbable as to have failed to occur in 2520 trials. While it is indeed doubtful that the lower boundary of the interval defined by the smallest recorded score is

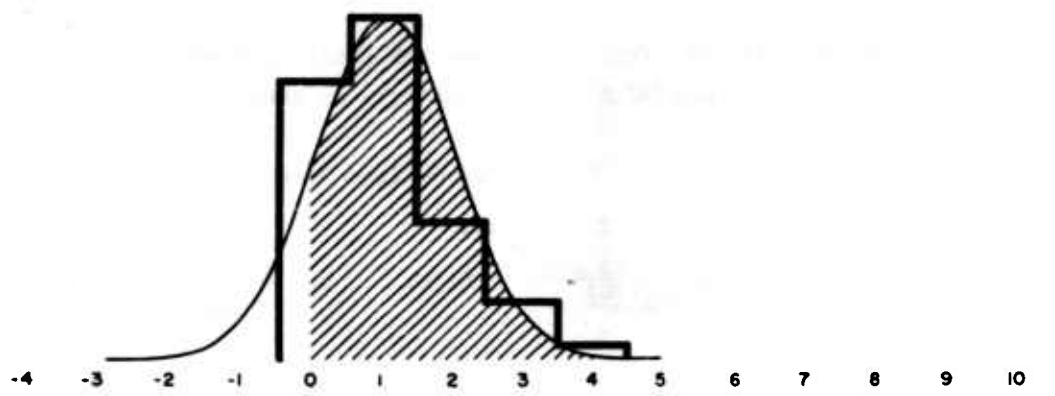

Figure 4. Error-Score Distribution (Histogram) and Normal Distribution with same Mean, Variance and "Area". Unshaded area of normal distribution covers impossible scores.

exactly the upper boundary for impossible scores, it would be difficult to deny that scores become impossible somewhere in this immediate vicinity. In any event, effects similar to, but less dramatic than, those shown in Tables I and II could have been shown using as "impossible" only negative time scores, about which definition there can be little dispute.

Distortions in probabilities induced by violations of assumptions are of critical importance at those probabilities calling for rejection of the null hypothesis. Usually the absolute discrepancy between "true" and "tabled" probabilities decreases with increasingly extreme significance levels. The relative, i.e., percent, error, however, increases rapidly and violations of assumptions generally produce their greatest relative distortions at the most extreme levels of significance. It is important in this context to note that, however small the tabled probability, if the true probability is zero and the tabled probability is not, the relative error is infinite.

TABLE I

THEORETICAL PROBABILITY, Pr(Z), FOR THE LEAST EXTREME NEGATIVE Z HAVING A TRUE PROBABILITY OF ZERO

Sample Size	Samples Drawn from Population of:			
	Time Scores		Error Scores	
N	$Z = \frac{5.5 - 15.1067}{\frac{12.1207}{\sqrt{N}}}$	Pr(Z)	$Z = \frac{0 - 1.0357}{\frac{.981331}{\sqrt{N}}}$	Pr(Z)
1	- .793	.2140	-1.055	.1457
2	- 1.121	.1311	-1.493	.0677
3	- 1.373	.0849	-1.828	.0338
4	- 1.585	.0565	-2.111	.0174
5	- 1.772	.0382	-2.360	.0091
6	- 1.941	.0261	-2.585	.0049
7	- 2.097	.0180	-2.792	.0026
8	- 2.242	.0125	-2.985	.0014
9	- 2.378	.0087	-3.166	.0008
10	- 2.506	.0061	-3.338	.0004
11	- 2.629	.0043	-3.500	.0002
12	- 2.746	.0030	-3.656	.0001
13	- 2.858	.0021	-3.805	.0001
14	- 2.966	.0015		
15	- 3.070	.0011		
16	- 3.170	.0008		
17	- 3.268	.0005		
18	- 3.363	.0004		
19	- 3.455	.0003		
20	- 3.545	.0002		

TABLE II

* LARGEST VALUE OF μ_o FOR WHICH IT IS IMPOSSIBLE TO REJECT THE FALSE NULL HYPOTHESIS $\mu \geq \mu_o$ USING THE ONE-TAILED Z TEST, $Z = \frac{\bar{X} - \mu_o}{\frac{\sigma}{\sqrt{N}}}$, at the α LEVEL OF SIGNIFICANCE

Sample Size	Samples Drawn from Population of: Time Scores (True μ = 15.1067)			Error Scores (True μ = 1.0357)		
N	α = .05	α = .01	α = .001	α = .05	α = .01	α = .001
1	25.44	33.69	42.95	1.614	2.283	3.032
2	19.60	25.44	31.98	1.141	1.614	2.144
3	17.01	21.78	27.12		1.318	1.751
4	15.47	19.60	24.23		1.141	1.516
5		18.11	22.25			1.356
6		17.01	20.79			1.238
7		16.16	19.66			1.146
8		15.47	18.74			1.072
9			17.98			
10			17.34			
11			16.79			
12			16.31			
13			15.89			
14			15.51			
15			15.17			

* Even for the last cell entry in each column the distorting effect of assumption violation is still large: when μ_o, the cell entry, equals μ, the true population mean, the tabled probability of rejection is α while the true probability of rejection is zero.

Since the entire analysis has been given in terms of the theoretical probability of an actually impossible score, it is an extremely conservative one, detecting only the most spectacular discrepancies between fact and theory. To the extent that it ignores discrepancies between theoretical and actual nonzero probabilities, it is therefore biased in favor of the philosophy which regards parametric tests as insensitive to violations of their assumptions. Obviously, the discrepancy between true and theoretical probabilities can still be extreme when both probabilities are greater than zero. For example, the lowest time score was a six which occurred with a frequency of one in 2520 trials. A mean of six for a sample of size N drawn, with replacement, from the time-score distribution can only be obtained, therefore, by drawing N sixes, and the true chance probability of doing so is $1/2520^N$.

The theoretical probability that $\bar{X} \leqq 6$ is found by referring $Z = \frac{6 - 15.1067}{\frac{12.1207}{\sqrt{N}}}$ to normal tables, using a one-tailed test. The following small table compares true and theoretical probabilities for those values of N which cause the theoretical probability to fall just within the .05, .01, and .001 levels of significance.

TABLE III

Theoretical and True Probabilities of Drawing a Certain "Possible" Sample

N	Theoretical $Pr(\bar{X} \leqq 6)$	True $Pr(\bar{X} \leqq 6)$	Percent Error
5	.04648	9.8×10^{-18}	4.7×10^{15}
10	.00875	9.7×10^{-35}	9.0×10^{31}
17	.00097	1.5×10^{-58}	6.5×10^{54}

It is clear from the table that the true and theoretical probabilities can be in extreme disharmony even though "possible" scores are used and that this may occur at realistic significance levels and reasonable sample sizes.

It is clear that violation of a single parametric assumption, that of normality, may give Type I and Type II errors a far different probability than the experimenter would believe to be the case on the basis of normal theory. It is equally clear that the degree to which probabilities of Type I and Type II errors are distorted by a given violation of assumptions is not a simple function of the "degree" of the violation, but is also greatly affected by the specific values of parameters, such as α and N, about which no assumptions have been made. If the experimenter is unaware of the nonnormality of his data, therefore, or if he believes that nonnormality is immaterial, he may err gravely in his statistical decisions, i.e., rejection or failure to reject, or in the "probabilities" which he associates with these decisions.

DISCUSSION

"Once making the assumptions, the mathematics is simple and exact and fascinatingly beautiful; and mathematicians will frankly say that it is our concern as researchers, not theirs, whether the assumptions are legitimate in the particular research situations with which we work. It happens that in most of the research in our field the assumptions are so far-fetched as to abort the results for careful work." So says Peters (66), speaking of Fisherian statistical tests. Thus, while some statisticians consider serious violations of assumptions to be rare, others regard them as commonplace. As implied in the passage quoted above, one reason for this is that the amount of assumption-violation typically encountered by research workers depends strongly upon the field in which research is conducted. The research worker's experiences naturally color his perspective. A distribution skewed to a degree commonly encountered by a psychologist might be regarded by him as mildly skewed and therefore constituting a moderate departure from the assumption of normality. The same degree of skewness may be seldom encountered by an agricultural statistician who may therefore pronounce it markedly skewed and a drastic departure from normality. A great deal of the

literature on assumption-violation was, in fact, contributed by agriculturally-oriented statisticians who attached "superlative" labels to degrees of assumption-violation commonly encountered by the writer, a psychologist. Thus the agricultural statistician's finding that moderate departure from the normality assumption does not appreciably distort significance levels may, when "translated", mean simply that when dealing with an unusually good fit to the normal distribution the psychologist may expect that the slight non-normality present will have no serious effect upon probabilities!

Besides varying widely in the degree of violation which they regard as "moderate" or "extreme", statisticians also differ vastly in the amount of error in probabilities which they deem a serious consequence of violation. The absence of commonly accepted and rigidly defined verbal classifications by which to categorize assumption violations and their effects makes a concise, quantitative summary of the literature very difficult. The task is further complicated by the fact that the effect of a violation of assumptions is not a simple function of the degree to which the assumption is violated. Instead, it is strongly influenced by factors which are not involved in the statement of the assumption and which would, in other circumstances, be irrelevant to the validity of the test: factors such as absolute sample size, relative sizes of treatment groups, size of significance level, and location of rejection region, i.e., whether in one or both tails. Furthermore if the same assumption is violated in qualitatively different ways by different treatment groups, the effect may be much more severe than if all groups had violated the assumption in the same way. For example, the effect of violating the normality assumption might be less if each of three groups were samples from the same positively skewed population than if one group were drawn from a positively skewed population and the other two from its negatively skewed mirror image with equal mean. Or again, if one treatment group is drawn from a rectangular, one from a triangular, and one from a skewed population, the distortion in probabilities may be greater than it would be for the worst case in which all three groups are drawn from the same one of these populations. Because of such interacting factors each investigation of the effects of assumption-violation is highly qualified by the conditions under which it was conducted and by the values of the parameters used.

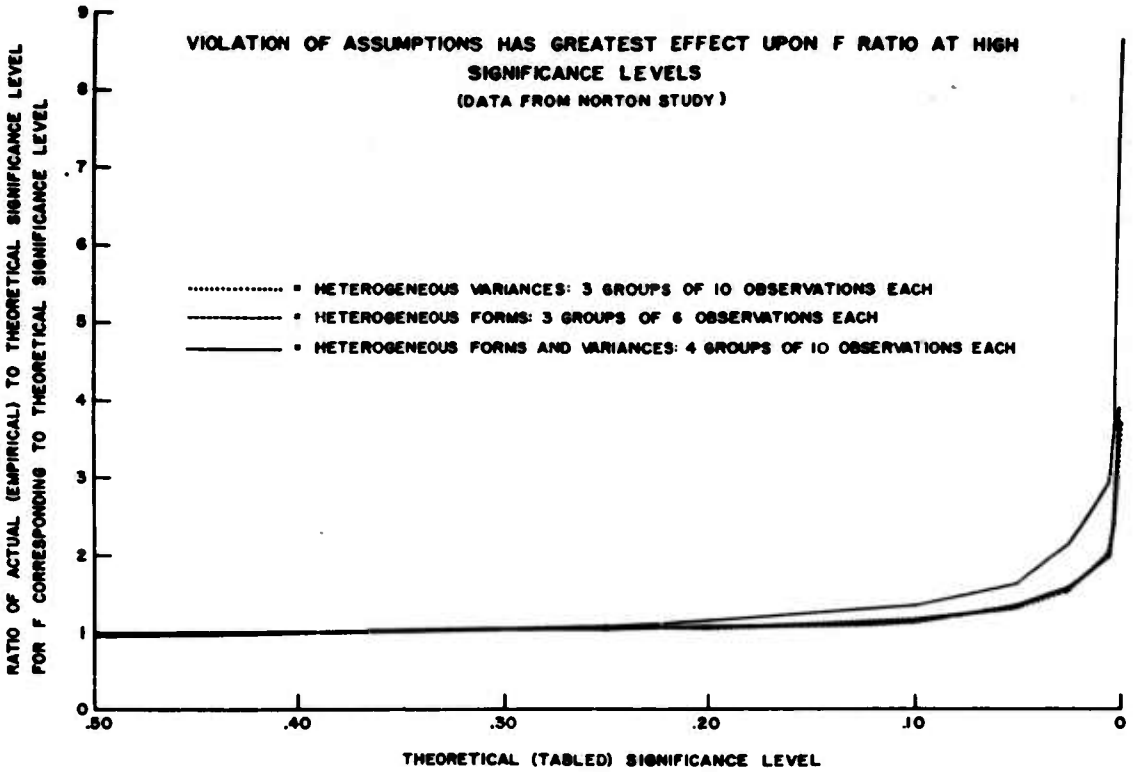

Figure 5. Data from Norton's study (59) show effects of several types of assumption-violation. At least 3000 F ratios were calculated for samples (a) ("heterogeneous variances") from normal populations with σ's of 5.0, 10.0, and 14.9, respectively, and means of 50.00, (b) ("heterogeneous forms") from a "normal" population limited in range to $\pm$ 2.5 σ's, an extreme-positive-skew and an extreme-negative-skew population, with σ's of 7.4, 6.2 and 6.2 and means of 50.00, 50.34 and 49.66, respectively, (c) ("heterogeneous forms and variances") normal, moderate-positive-skew, extreme-positive-skew, and L-shaped populations with σ's of 14.90, 10.02, 6.20, and 2.24 and means of 50.00, 50.27, 50.34 and 50.00, respectively. Curves in figure show ratio of empirical to "normal theory" cumulative F distributions, cumulated to various standard significance levels, for each of three assumption-violating sampling situations.

This high specificity effectively prevents quantitative generalization of the diverse results reported in the literature. However, certain qualitative generalizations appear to be firmly supported. Under a given violation of assumptions the resulting error in probability level tends to increase as sample size decreases or as significance levels become more extreme. Furthermore, the probability error would appear to be reduced by following a sort of symmetry principle. Under many circumstances the distortion increases as groups become more unequal in size, or as the sampled populations become more dissimilar in form, and the distortion is frequently greater when a one-tailed test is used than when the test is two-tailed. Finally, it appears that combinations of violations, or of aggrevating conditions such as those listed above, tend to have an aggregate effect which exceeds the sum of their separate effects.

For quantitative results the reader is referred to the literature. There he will find a diversity of situations regarded by their investigators as conducive to appreciable error. Even the most sanguine of investigators tend (in the body of the report) to draw attention to one or more of such "serious" conditions of violation. Because of the great variation in the amount of error which they regard as "serious", however, nothing would be gained by reporting their opinions. In most of these studies the absolute difference between true and theoretical probabilities amounts to only a few hundredths or even thousandths. It is important, however, to recall that (except at the very smallest values of N) the same can be said of the investigation reported herein in which a small absolute error in probabilities had a tremendous effect upon the power of the test.

When their assumptions are in any way violated, parametric tests are practically certain to yield inexact probabilities. In this case they are approximate tests which, if the approximation is good, may identify the general neighborhood in which the true probability lies but which do not specify it precisely. Frequently, perhaps usually, this is all that is desired. It is important to note, however, that certain virtues enjoyed by parametric tests when all assumptions are met can no longer be claimed for them when violations of assumptions reduce them to approximate tests. The

TABLE IV

Effect of f_e, p, n and α upon the accuracy of Chi Square, χ^2, and Chi Square with Yates' Correction, χ^2_y, in Testing the Significance of the Observed Frequency of a Binomial Event

(f_e = Expected Frequency, p = Probability that Event will Occur in a Single Trial, n = Sample Size, α = Significance Level)

			Tabulated $Pr(\chi^2)$ Corresponding to True α of:			Tabulated $Pr(\chi^2_y)$ Corresponding to True α of:		
$f_e = np$	p	n	.05	.01	.001	.05	.01	.001
50	.50	100	.0395	.0077	.0008	.0508	.0104	.0011
50	.25	200	.0414	.0079	.0007	.0505	.0102	.0010
50	.10	500	.0422	.0081	.0007	.0506	.0100	.0009
50	.05	1000	.0426	.0080	.0007	.0506	.0099	.0009
20	.50	40	.0349	.0069	.0007	.0518	.0112	.0013
20	.25	80	.0369	.0070	.0006	.0517	.0105	.0010
20	.10	200	.0386	.0069	.0005	.0514	.0100	.0008
20	.05	400	.0391	.0071	.0004	.0514	.0099	.0007
10	.50	20	.0306	.0068	.0009	.0534	.0141	.0020
10	.25	40	.0337	.0062	.0004	.0528	.0109	.0009
10	.10	100	.0345	.0055	.0003	.0532	.0094	.0006
10	.05	200	.0352	.0056	.0002	.0532	.0094	.0004
10	.01	1000	.0358	.0057	.0002	.0533	.0093	.0004
5	.50	10	.0266	.0080	---	.0625	.0226	---
5	.25	20	.0288	.0049	.0003	.0591	.0114	.0008
5	.10	50	.0310	.0050	.0002	.0571	.0102	.0004
5	.05	100	.0319	.0031	.0001	.0571	.0067	.0002
5	.01	500	.0327	.0032	.0001	.0572	.0067	.0002
2	.50	4	---	---	---	---	---	---
2	.25	8	.0262	.0026	.0001	.0755	.0101	.0006
2	.10	20	.0294	.0008	.0000	.0723	.0033	.0001
2	.05	40	.0307	.0009	.0000	.0726	.0032	.0001
2	.01	200	.0140	.0009	.0000	.0400	.0032	.0000
1	.50	2	---	---	---	---	---	---
1	.25	4	.0068	.0014	---	.0497	.0099	---
1	.10	10	.0061	.0002	.0000	.0329	.0013	.0000
1	.05	20	.0065	.0002	.0000	.0325	.0013	.0000
1	.01	100	.0070	.0002	.0000	.0326	.0013	.0000

Note: The Chi Square test assumes that, for each cell, observed frequencies are normally distributed about their expected frequency. This assumption is fully met only when n is infinite. As shown in above table, the effect of violation is a complex function of several parameters, not just of f_e. See Appendix for derivation of table.

writer has frequently heard the belief expressed that even when its assumptions are violated a parametric test is still preferable to a distribution-free test, all of whose assumptions are met, because the parametric test is "more efficient". The efficiency of test A relative to test B is generally defined as the ratio b/a where a is the number of observations upon which test A must be based in order to equal the power of test B based upon b observations. Both tests must be applied to the same data under the same conditions, e.g., using the same significance level. Almost all efficiency figures for distribution-free, relative to parametric, tests have been calculated for the case where both tests are applied to data meeting all of the assumptions of the parametric comparison statistic. Practically all efficiency figures for distribution-free tests relative to parametric tests are therefore inapplicable when the assumptions of the parametric test are false. The contention that, when its assumptions are violated, a parametric test is still to be preferred to a distribution-free test because it is "more efficient" is therefore a monumental non sequitur. The point is not at all academic. Table II shows that violations of a test's assumptions may be attended by profound changes in its power. And since the efficiency of a test is a function of its power, it is clear that the nominal efficiency may be gravely in error when the test is used in violation of its assumptions. In the cases shown in Table II, violation of the assumption of normality rendered a parametric test completely powerless over a fairly broad area of applications. However, violation of this parametric assumption would have no effect upon the power of a distribution-free test whose assumptions had been met. The efficiency of the distribution-free test, relative to the parametric test, would therefore, for the cases covered by Table II, be infinite.

As has been seen, when their assumptions are violated parametric tests tend to err the most at small sample sizes and extreme levels of significance, the relative error diminishing rapidly with increasing N or with increasing α. Furthermore at very small values of N only a priori knowledge can constitute legitimate evidence that the assumptions are satisfied; when applied to very small samples, tests for the validity of the assumptions are unlikely to detect any but the most drastic violations. As generally applied

in the behavioral sciences, therefore, parametric tests are of extremely dubious validity when based on very small samples or when employing extreme levels of significance. These, however, are precisely the conditions under which distribution-free tests are most accurate and most "efficient". When sample size is small, probabilities for most distribution-free tests are calculated from exact combinatorial formulae. Beginning at moderate values of N, however, (usually from an N of from 10 to 20 upwards) exact calculations involve a prohibitive amount of labor for all but the most extreme levels of probability, and approximate probabilities are obtained using asymptotic, i.e., inexact, formulae. Frequently the approximation is a very close one; however, for some tests it is rather poor at the smallest values of N at which asymptotic formulae are used to calculate probabilities. The efficiencies of distribution-free tests relative to their parametric counterparts range from zero to one in the asymptotic case of infinitely large samples. Their efficiencies increase, however, with decreasing sample size, and generally reach a value surprisingly close to one when sample size becomes ten or less (apparently approaching one or a value very close to it as sample size approaches zero). The conclusion is obvious. When sample size is very small, unless an experimenter knows from prior considerations that all assumptions are satisfied, use of a parametric rather than a distribution-free test involves great risk and the prospect of little or no gain. The writer would suggest use of the following rule of thumb in the absence of such prior knowledge that parametric assumptions are satisfied:

> Use a distribution-free test (a) when $N \leqq 10$, whatever the value selected for α, (b) when $N \leqq 20$ and $\alpha < .05$, (c) when $N \leqq 30$ and $\alpha < .01$. In other cases consider such indications of the type and extent of violation as are available, weigh the nominal efficiencies of the respective tests at the sample size contemplated (bearing in mind that nominal and true efficiencies may differ greatly if assumptions are violated), and be guided by the perspective afforded by a review of the literature on assumption-violation.

Should a parametric test be elected, the experiment should be designed so as to mitigate the effects of assumption-violation. Sample size should be

no smaller and significance level no more extreme than necessary. Various devices may be employed to reduce the effects of violation. For example, by applying a t-test to matched pairs rather than to unmatched data, one deals with a population distribution which is likely to be at least symmetrical, if not normal, if the null hypothesis is true; and, since only one variance, that of the difference-scores, is involved, the assumption of homogeneity of variance is "eliminated".

SUMMARY AND CONCLUSIONS

The unqualified generalization that parametric tests are insensitive to violation of their assumptions is a dangerous fallacy apparently attributable to wishful thinking, biassed methods and overgeneralization of results. The fact is that violation of parametric assumptions may have negligible or serious effect upon probabilities depending upon a multiplicity of factors. Many of these are not involved in the statement of the assumptions but interact dramatically with whatever violation exists. For a given violation of assumptions the resulting distortion in probabilities tends to increase with diminishing sample size and with diminishing significance level. Frequently it increases with inequality of group sizes, dissimilarity of sampled populations, and is greater for one-tailed than for two-tailed tests. A violation of assumptions which causes a small absolute error in probability levels may produce a large relative error and a great change in the power of the test. Therefore, when parametric assumptions are violated a distribution-free statistic may be far more efficient than its nominally more efficient parametric counterpart.

BIBLIOGRAPHY

1. Adler, F. Yates' correction and the statisticians. J. Amer. Statist. Ass., 1951, 46, 490-501.

2. Bartlett, M. S. The effect of non-normality on the t distribution. Proc. Cambr. Phil. Soc., 1935, 31, 223-231.

3. Bartlett, M. S. The use of transformations. Biometrics, 1947, 3, 39-52.

4. Box, G. E. P. Non-normality and tests on variances. Biometrika, 1953, 40, 318-335.

5. Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. Ann. Math. Statist., 1954, 25, 290-302.

6. Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effects of inequality of variance and of correlation between errors in the two-way classification. Ann. Math. Statist., 1954, 25, 484-498.

7. Bradley, J. V. Studies in research methodology, I: Compatibility of psychological measurements with parametric assumptions. WADC Technical Report No. 58-574 (1), 1958, In Press.

8. Bradley, R. A. Corrections for nonnormality in the use of the two-sample t- and F- tests at high significance levels. Ann. Math. Statist., 1952, 23, 103-113.

9. Bradley, R. A. The distribution of the t and F statistics for a class of non-normal populations. Virginia J. Sci., 1952, 3, 1-32.

10. Camp, B. H. The effect on a distribution function of small changes in the population function. Ann. Math. Statist., 1946, 17, 226-231.

11. Chesire, Leone; Oldis, Elena, & Pearson, E. S. Further experiments on the sampling distribution of the correlation coefficient. J. Amer. Statist. Assn., 1932, 28, 121-128.

12. Chung, K. L. The approximate distribution of Student's statistic. Ann. Math. Statist., 1946, 17, 447-465.

13. Church, A. E. R. On the means and squared standard-deviations of small samples from any population. Biometrika, 1926, 18, 321-394.

14. Cochran, W. G. Some consequences when the assumptions for the analysis of variance are not satisfied. Biometrics, 1947, 3, 22-38.

15. Cochran, W. G. The X^2 correction for continuity. Iowa State College J. Sci., 1942, 16, 421-436.

16. Cochran, W. G. The X^2 distribution for the binomial and Poisson series, with small expectations. Annals of Eugenics, 1936, 7, 207-217.

17. Cochran, W. G. The X^2 test of goodness of fit. Ann. Math. Statist., 1952, 23, 315-345.

18. Cochran, W. G. The statistical analysis of field counts of diseased plants. J. Roy. Statist. Soc.(B), 1936, 3, 49-67.

19. Cochran, W. G., & Cox, Gertrude. Experimental designs. New York: Wiley, 1950, 83-84.

20. Crow, E. L. Some cases in which Yates' correction should not be applied. J. Amer. Statist. Ass., 1952, 47, 303-304.

21. David, Florence N., & Johnson, N. L. The effect of non-normality on the power function of the F-test in the analysis of variance. Biometrika, 1951, 38, 43-57.

22. Dixon, W. J., & Massey, F. J. Introduction to statistical analysis. New York: McGraw-Hill, 1951, pp. 99, 103, 106, 107, 127, 133, 138, 139, 141, 162.

23. Dunlap, Hilda F. An empirical determination of the distribution of means, standard deviations and correlation coefficients drawn from rectangular populations. Ann. Math. Statist., 1931, 2, 66-81.

24. Edwards, A. L. Experimental design in psychological research. New York: Rinehart, 1950, pp. 195-207.

25. Edwards, A. On "The use and misuse of the chi-square test" - the case of the 2 x 2 contingency table. Psychol. Bull., 1950, 47, 341-346.

26. Eisenhart, C. The assumptions underlying the analysis of variance. Biometrics, 1947, 3, 1-21.

27. Finch, D. J. The effect of non-normality on the z-test, when used to compare the variances of two populations. Biometrika, 1950, 37, 186-189.

28. Fisher, R. A. On a property connecting the X^2 measure of discrepancy with the method of maximum likelihood. Atti del Congresso Internazionale dei Mathematici, Bologna, 1928, 6, 95-100.

29. Fisher, R. A. "The coefficient of racial likeness" and the future of craniometry. J. Roy. Anthrop. Inst. Great Britian and Ireland, 1936, 66, 57-63.

30. Fisher, R. A. The conditions under which X^2 measures the discrepancy between observation and hypothesis. J. Roy. Statist. Soc., 1924, 87, 442-450.

31. Fisher, R. A. The design of experiments. New York: Hafner, 1953, 43-47.

32. Freeman, G. H., & Halton, J. H. Note on an exact treatment of contingency, goodness of fit and other problems of significance. Biometrika, 1951, 38, 141-149.

33. Fry, T. C. The X^2-test of significance. J. Amer. Statist. Ass., 1938, 33, 513-525.

34. Gayen, A. K. Significance of difference between the means of two non-normal samples. Biometrika, 1950, 37, 399-408.

35. Gayen, A. K. The distribution of 'Student's' t in random samples of any size drawn from non-normal universes. Biometrika, 1949, 36, 353-369.

36. Geary, R. C. Testing for normality. Biometrika, 1947, 34, 209-242.

37. Geary, R. C. The distribution of 'Student's' ratio for non-normal samples. J. Roy. Statist. Soc. (B), 1936, 3, 178-184.

38. Geisser, S. A note on the normal distribution. Ann. Math. Statist., 1956, 27, 858-859.

39. Greenhood, E. R. Detailed proof of the chi-square test of goodness of fit. Cambridge, Mass.: Harvard Univ. Press, 1940.

40. Grownow, D. G. C. Non-normality in two-sample t-tests. Biometrika, 1953, 40, 222-225.

41. Gumbel, E. J. On the reliability of the classical chi-square test. Ann. Math. Statist., 1943, 14, 253-263.

42. Hastings, C., et al. Low moments for small samples: a comparative study of order statistics. Ann. Math. Statist., 1947, 18, 413-426.

43. Hey, G. B. A new method of experimental sampling illustrated on certain non-normal populations. Biometrika, 1938, 30, 68-80.

44. Hill, I. D. The distribution of the regression coefficient in samples from a non-normal population. Biometrika, 1954, 41, 548-552.

45. Horsnell, G. The effect of unequal group variances on the F-test for the homogeneity of group means. Biometrika, 1953, 40, 128-136.

46. Hotelling, H. Effects of non-normality at high significance levels (abstract). Ann. Math. Statist., 1947, 18, 608-609.

47. Laderman, J. The distribution of "Student's" ratio for samples of two items drawn from non-normal universes. Ann. Math. Statist., 1939, 10, 376-379.

48. Lancaster, H. O. Statistical control of counting experiments. Biometrika, 1952, 39, 419-422.

49. Lancaster, H. O. The exact partition of X^2 and its application to the problem of the pooling of small expectations. Biometrika, 1950, 37, 267-270.

50. Lewis, D., & Burke, C. J. Further discussion of the use and misuse of the chi-square test. Psychol. Bull., 1950, 47, 347-355.

51. Lewis, D., & Burke, C. J. The use and misuse of the chi-square test. Psychol. Bull., 1949, 46, 433-489.

52. Lindquist, E. F. Design and analysis of experiments in psychology and education. New York: Houghton Mifflin, 1953, 72-90.

53. Lukacs, E. A characterization of the normal distribution. Ann. Math. Statist., 1942, 13, 91-93.

54. McNemar, Q. Psychological statistics. New York: Wiley, 1949, pp. 103, 108, 113, 120, 167, 170, 173, 174, 178, 179, 182, 198, 216, 223, 225, 231, 235, 241, 249, 324.

55. Nair, A. N. K. Distribution of Student's 't' and the correlation coefficient in samples from non-normal populations. Sankhyā, 1941, 5, 383-400.

56. National Bureau of Standards. Tables of normal probability functions. Washington, D. C.: U. S. Govt. Printing Office, 1953.

57. Neyman, J. On the correlation of the mean and variance in samples drawn from an "infinite" population. Biometrika, 1926, 18, 401-413.

58. Neyman, J., & Pearson, E. S. Further notes on the X^2 distribution. Biometrika, 1930, 22, 298-305.

59. Norton, D. W. An empirical investigation of the effects of nonnormality and heterogeneity upon the F-test of analysis of variance. Ph.D. Dissertation, Department of Education, State University of Iowa, August 1952.

60. Pastore, N. Some comments on "The use and misuse of the chi-square test". Psychol. Bull., 1950, 47, 338-340.

61. Pearson, E. S. Some notes on sampling tests with two variables. Biometrika, 1929, 21, 337-360.

62. Pearson, E. S. The analysis of variance in cases of non-normal variation. Biometrika, 1931, 23, 114-133.

63. Pearson, E. S. The test of significance for the correlation coefficient. J. Amer. Statist. Ass., 1931, 26, 128-134.

64. Pearson, E. S., & Adyanthāya, N. K. The distribution of frequency constants in small samples from non-normal symmetrical and skew populations. Biometrika, 1929, 21, 259-286.

65. Perlo, V. On the distribution of Student's ratio for samples of three drawn from a rectangular distribution, Biometrika, 1933, 25, 203-204.

66. Peters, C. C. Misuses of the Fisher statistics. J. Educ. Res., 1943, 36, 546-549.

67. Peters, C. C. The misuse of chi-square - a reply to Lewis and Burke. Psychol. Bull., 1950, 47, 331-337.

68. Keitz, H. L. On the distribution of the "Student" ratio for small samples from certain non-normal populations. Ann. Math. Statist., 1939, 10, 265-274.

69. Rider, P. R. On small samples from certain non-normal universes. Ann. Math. Statist., 1931, 2, 48-65.

70. Rider, P. R. On the distribution of the correlation coefficient in small samples. Biometrika, 1932, 24, 382-403.

71. Rider, P. R. On the distribution of the ratio of mean to standard deviation in small samples from non-normal universes. Biometrika, 1929, 21, 124-141.

72. Shewhart, W. A., & Winters, F. W. Small samples - new experimental results. J. Amer. Statist. Ass., 1928, 23, 144-153.

73. Staff of the Computation Laboratory, Harvard University. Tables of the cumulative binomial probability distribution. Cambridge, Mass.: Harvard University Press, 1955.

74. Student, The probable error of a mean. Biometrika, 1908, 6, 1-25.

75. Sukhatme, P. V. On the distribution of X^2 in samples of the Poisson series. J. Roy. Statist. Soc. (B), 1938, 5, 75-79.

76. Tukey, J. W. Some elementary problems of importance to small sample practice. Human Biol., 1948, 20, 205-214.

77. van der Vaart, H. R. On certain statistical methods used in biology with special reference to Husson's paper on cricetus cricetus canescens nehring. Proc. Koninkl. Nederl. Akad. Wetensch., Series C, 1953, 56, 631-638.

78. Walsh, J. E. A large sample t-statistic which is insensitive to non-randomness. J. Amer. Statist. Ass., 1951, 46, 79-88.

79. Welsh, B. L. On tests for homogeneity. Biometrika, 1938, 30, 149-158.

80. Williams, C. A. On the choice of the number and width of classes for the chi-square test of goodness of fit. J. Amer. Statist. Ass., 1950, 45, 77-86.

81. Yates, F. Contingency tables involving small numbers and the X^2 test. J. Roy. Statist. Soc. (B), 1934, 1, 217-235.

APPENDIX

METHOD USED TO OBTAIN TABLE IV

Let r be the observed number of occurrences, in n binomial trials, of an event whose constant probability of occurrence on a single trial is p. The mean, or expected frequency, is therefore np. In order that exact probabilities may be obtained from the binomial tables (73), choose n and p in such a way that p is in hundredths and both n and np are integers and let d be the smallest integer such that $Pr\ (r \leqq np - d) + Pr\ (r \geqq np + d) \leqq \alpha$, where α is one of the standard significance levels, .05, .01 or .001. The exact, true probability of a deviation of d or greater between observed and expected frequencies is therefore given by $Pr\ (r \leqq np - d) + Pr\ (r \geqq np + d)$, and this is the largest such probability less than α.

Using the chi-square, rather than the binomial, test to obtain the probability of a deviation of d or greater, the observed and expected frequencies of occurrence, fo_1 and fe_1, will equal r and np respectively, the corresponding frequencies of nonoccurrence, fo_2 and fe_2, will equal n - r and n - np respectively, and chi-square will have a single degree of freedom. Applying the chi-square test in the usual manner, a probability will be obtained which differs from that obtained by the exact binomial method. In order to obtain the chi-square probability corresponding to a true probability of α, the chi-square probability must be multiplied by the ratio of α to the true binomial probability, $Pr\ (r \leqq np - d) + Pr\ (r \geqq np + d)$; the product then becomes a cell entry in Table IV.

Actually, since elaborate chi-square tables were not at hand, "chi square" probabilities were obtained from normal tables (56) by making use of the fact that, when based upon a single degree of freedom, chi is normally distributed. In the binomial case, using Yates' correction,

$$\chi^2_y = \frac{(\,|\,fo_1 - fe_1\,|\,-1/2)^2}{fe_1} + \frac{(\,|\,fo_2 - fe_2\,|\,-1/2)^2}{fe_2}$$

$$= \frac{(\,|\,r - np\,|\,-1/2)^2}{np} + \frac{(\,|\,(n-r) - (n-np)\,|\,-1/2)^2}{n - np}$$

$$= \frac{(|r - np| - 1/2)^2}{np} + \frac{(|r - np| - 1/2)^2}{n - np}$$

$$= \frac{n(|r - np| - 1/2)^2}{np(n - np)}$$

$$X_y = \frac{|r - np| - 1/2}{\sqrt{np(1 - p)}}$$

Thus $X_y = \frac{d - 1/2}{\sqrt{np(1-p)}}$ and, by similar derivation, $X = \frac{d}{\sqrt{np(1-p)}}$. These values were referred to normal tables (56) to obtain, to the number of decimal places needed, the probabilities which elaborate chi square tables would have given for X_y^2 and X^2.

The chi square test assumes that, in each cell, observed frequencies are normally distributed about their expected frequency. This is illustrated in the above formula for chi where the expression on the right hand side of the equation is obviously the normal approximation to the binomial. Since a normal distribution extends to infinity on both sides of the mean and since observed frequencies cannot be less than zero, the normality assumption is "met" only when the mean, np, is infinitely large. This can occur only if the sample size, n, is infinite. Thus, although sometimes called "nonparametric", the normality assumption and a notorious susceptibility to misapplication give the chi square test much more in common with parametric than with distribution-free tests.